Linking Semantically Enabled Online Community Sites

Article · September 2004

CITATIONS
5

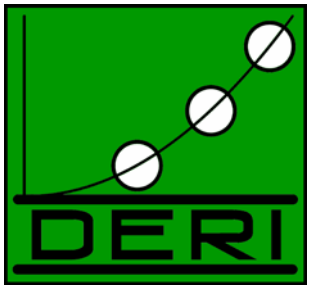
READS
92

4 authors, including:

Andreas Harth
Friedrich-Alexander-University of Erlangen-Nürnberg
178 PUBLICATIONS **4,574** CITATIONS
SEE PROFILE

John G Breslin
University of Galway
446 PUBLICATIONS **7,966** CITATIONS
SEE PROFILE

Ina O'Murchu
17 PUBLICATIONS **166** CITATIONS
SEE PROFILE



Linking Semantically Enabled Online Community Sites

Andreas Harth, John G. Breslin, Ina O'Murchu,

Stefan Decker

DERI Technical Report 2004-08-09

August 2004

DERI Galway
University Road
Galway
IRELAND
www.deri.ie

DERI Innsbruck
Technikerstrasse 13
A-6020 Innsbruck
AUSTRIA
www.deri.at

Linking Semantically-Enabled Online Community Sites

Andreas Harth, John G. Breslin, Ina O'Murchu, Stefan Decker
Digital Enterprise Research Institute, National University of Ireland, Galway,
University Road, Galway, Ireland
{andreas.harth, john.breslin, ina.omurchu, stefan.decker}@deri.org

Abstract

Online community sites have replaced the traditional means of keeping a community informed via libraries and publishing. At present, online communities are islands that are not interlinked. We describe different types of online communities and tools that are currently used to build and support such communities. Ontologies and semantic web technologies offer an upgrade path to providing more complex services. Fusing information and inferring links between the various applications and types of information provides relevant insights that make the available information on the Internet more valuable. We present the SIOC ontology which combines terms from vocabularies that already exist with new terms needed to describe the relationships between concepts in the realm of online community sites.

1 Introduction

Online community sites have replaced the traditional means of keeping a community informed via libraries and publishing. These sites allow improved communication and interactive contact within a community, by providing an online collaboration space for certain interest-related or localized information. Members can find and contribute relevant shared information and ideas to others within the site. Community sites are equally suited for both profit and non-profit purposes (professional or social) [O'Murchu *et al.*, 2004].

Semantic Web technologies can be used to enrich community sites and to make the underlying information available to both humans and software agents. Most community sites process and share information amongst their members through a personalized central point, and search queries for information are usually keyword based. The current web technologies employed by community sites are a serious limitation in making information accessible to their users in an efficient manner [Lara *et al.*, 2004]. In our paper we will explain how the use of Semantic Web technologies can enable community sites to become more efficient at the task of sharing and searching for information relevant to that community.

At present, online communities are islands that are not interlinked. Search facilities are also limited to syntactic matching, e.g., by message keyword on bulletin boards. Many communities can exist that are discussing complementary topics. For example, a forum message on community site A may be related to an email message on mailing list B, but a forum search will not represent this. Once there exists enough sites that have richer query facilities, then these different sites can be interlinked (in the example, the message on site A can then be related to the message on site B). Also, once a user has an account at site A, then site B could pull that user information from site A and would not need to maintain their own user accounts database[1].

Other benefits of having uniform access to community-related data in a semi-structured format as RDF include: applying reasoning facilities (such as those provided by OWL-S time [Pan and Hobbs, 2004]) to make use of events descriptions), visualization of scheduling information of individuals or groups of people, integration of search facilities over all tools in an interlinked set of community sites, and a representation of where people related to a certain topic are located geographically. Interlinking these various parts into a coherent representation enables more sophisticated applications and therefore results in more efficient information dissemination in communities.

For example, a user is searching for information on installing broadband on a Linux-based PC in their house in Galway. There is a post discussing local internet service providers on a bulletin board dedicated to Galway that references both a Usenet post comparing different broadband modems and a mailing list post detailing how to install broadband on Linux. Previously the user would have to traverse three sites to find the relevant information. However, depending on whether remote querying is possible or if external data is being warehoused (e.g., warehousing of relevant posts, people, events or forums one level or two levels away from a local community), a search for broadband on the Galway bulletin board will also yield the relevant text from the interlinked Usenet and mailing list community posts.

We will begin by describing what tools online communities currently use (such as bulletin boards, wikis, weblogs, and so on), and we will then show how these tools are already integrated in the current HTML web or the Semantic Web if that is the case. Then, we will describe our SIOC (Semantically-Interlinked Online Community) ontology for community sites that can accommodate information from community tools by mixing and matching already existing ontologies and extending them where appropriate. Finally, we discuss how the data these tools provide can be integrated into a semantically-enabled online community site, therefore linking up many disparate communities on the Semantic Web.

2 Existing Community Tools

Different roles exist in online communities: information providers that publish information; information consumers that consume information; and infrastructure providers that provide the infrastructure needed for publishing and exchanging information. To be able to access information on the Semantic Web in RDF, infrastructure providers such as webmasters or mailing list administrators have to open up their databases for semantic web access technology. The core technologies are available, but now the most important part is social: to convince people to adopt these technologies [Hendler, 2004]. We will show in this section what tools are already available, and demonstrate in subsequent sections how the tools can be integrated and linked up on the Semantic Web, therefore yielding some practical examples on how administrators can enable their sites for the Semantic Web.

Different types of communities will have diverse requirements [Wellmann and Gulia, 1999]. Informal social communities may only require the exchange of messages between users, whereas professional communities will focus more on the exchange of documents using protocols such as FTP or CVS. Professional communities have the requirement to coordinate calendars by exchanging event information from calendaring tools. We defer the topic of information sharing in professional communities to future work, and will focus this paper to describe requirements for informal online communities that are prevalent on the Internet.

The following subsections will briefly introduce the tool or protocol used and then will describe how a

[1] http://www.phpbb.com/phpBB/viewtopic.php?t=197635

particular tool fits into the current web infrastructure or how a tool is already integrated with other technologies. The most commonly used approach for integrating community tools into the current web is for the tools to use a HTML user interface that will provide access to the required data. Some of the newer tools already publish their data in RDF, and are therefore already enabled for the Semantic Web.

2.1 Mailing Lists

Email is still the most prevalent asynchronous one-to-many communication medium on the Internet. Mailing lists provide a quick method to set up communication features for an online community. Mailing lists were also one of the first methods used to set up and support a closed-group online community. Unfortunately, email and mailing lists can be subject to abuse (e.g., mail bombs, spam, or other unsolicited mail).

Although email's main transport protocols are SMTP, POP3, and IMAP4, and the format is text-based (RFC 822[2]), the contents of mailing lists are also being made available on the Web in HTML format. For example, Yahoo! Groups (formerly eGroups) allows the creation of private or public community mailing lists, with messages browsable via the Web and/or sent via individual or digest-type emails. Archives of mailing lists hosted on individual servers are often made available online in HTML, using tools such as GNU Mailman or MHonArc. Some mailing lists, such as DBWorld[3], already have message headers defined to include annotations in semi-structured format, e.g., metadata descriptions about calls for papers.

2.2 Usenet

Newsgroups, a collection of server-distributed discussion groups, are one of the oldest community-building primitives in existence. The most popular newsgroup system is Usenet, which is used to exchange knowledge and ask for help on a broad range of topics. On Usenet, there is no access control: everybody who is able to install the required tools can participate. Every message has a unique message id as specified in RFC 1036[4] similar to email. Usenet is based on the NNTP protocol which transmits messages over the Internet.

Methods for integrating newsgroups into the HTML web include systems such as the DNewsWeb or WebNews server system, which provide a proxy connection to an NNTP server via a web server. Also, a number of NNTP gateway and integration add-ons have been written for popular bulletin board systems so that messages from both bulletin board forums and Usenet can be browsed through a single community site. These systems usually import previous Usenet messages periodically into the same SQL database as the board, and new messages posted locally are exported and sent to the news server.

[2] http://www.faqs.org/rfcs/rfc822.html
[3] http://www.cs.wisc.edu/dbworld/
[4] http://www.faqs.org/rfcs/rfc1036.html

2.3 Bulletin Boards

The bulletin board has been a popular feature of internet-based communication since the early days of mailing lists and Usenet newsgroups. One of the quickest and most common ways of establishing an online community for those with like-minded interests is via the creation of a bulletin board. A bulletin board normally contains a set of forums classified into categories, and may also integrate event meeting calendars.

Bulletin boards have evolved beyond the traditional admin-maintained structure into one where forums and categories can be created once a critical mass of user support has been received. Most forums on community sites employ some threaded display methods, where topics are initialized by a certain user and replied to by others. The moderator of a forum has the responsibility for pruning undesirable threads and banning unwanted users from the forum. An administration discussion forum can raise useful suggestions or bug reports that can increase the usability of the underlying software.

Bulletin boards are a thriving part of the current HTML web. Posts on a bulletin board can be referenced via a URI. Some popular bulletin board systems include vBulletin, phpBB, Invision Board, and the ezboard forum hosting service. Websites of open source projects such as those hosted on sourceforge.net include forum functionality to enable discussions between project members and software users.

2.4 Chat

What email is to asynchronous communication on the Internet, chat is to synchronous communication. Chat on the Internet comes in various flavors: Internet Relay Chat (IRC) as specified in RFC 1459[5], Instant Messaging (IM), and web-based chats. IRC has long been used by communities to host real-time discussions of various topics, divided into chat rooms or channels for each topic. IM enables a speedy and efficient exchange of text messages in real time, usually between two people, however one-to-many IM sessions are possible. Although chat is primarily used to send and receive text messages, file transfer is also possible (e.g., using DCC on IRC).

Multi-protocol IM clients such as Trillian or Gaim can integrate the various IM networks and can connect to IRC. IRC chat is usually integrated into community sites by means of Java applets, such as PJIRC or JPilot. Automated bots can log into IRC channels and record real-time discussions for archival or statistical purposes and publish the logs in either text, HTML, XML or RDF. They can also perform more complex actions like interactive quizzes or weather updates for a particular area via restricted natural language input.

2.5 Weblogs

Weblogs are websites that are updated habitually by their creators, who provide brief news entries that are presented in chronological order. A weblog can be

[5] http://www.faqs.org/rfcs/rfc1459.html

produced by anyone with no previous knowledge of programming or HTML editing, using a simple web-based interface. RDF Site Summary[6] (RSS 1.0), an RDF-based format that can be used to exchange weblog entries, enables the syndication of one blog's content into another blog (or into a news reader or aggregator). Atom is a newly specified XML-based format to overcome the shortcomings of RSS. Blogrolls are used by bloggers, who tend to publish a list of the blogs they read on a regular basis along the side of their own blog and therefore highlighting other sites of interest to the reader.

There are several popular web logging software publishing tools available at present such as Movable Type and LiveJournal. Movable Type is a publishing system which installs on web servers to enable individuals or organizations to manage and update weblogs, journals, and other frequently-updated website content. LiveJournal is another simple blogging tool based on open source software. Many community sites have begun to incorporate blogging features, e.g., vBJournal is an integration of bulletin board and weblog technologies. Collaborative weblogs are more akin to bulletin board discussions, where more than one person in either a public or closed group can contribute to a shared weblog.

An IRC bot can also be used to create a collaborative weblog or "scratchpad" from an IRC chat channel where all members of the channel provide content. An example is the Daily Chump bot on the IRC channel #rdfig.

2.6 Wikis

A wiki, a collaboratively edited website, allows a community open read and write access to a database of pages on a site, even if a user is not the originator of the material being edited. Users can create new pages or change existing pages easily via a web-based interface. The original WikiWikiWeb did not have any access control, and anybody could participate in the live editing of pages. This flexibility can either be successful in a busy community or disastrous in an indifferent community (where anonymous users can vandalize or make unwanted changes to a wiki set). Many wikis now feature a version control system so that rollback to a previous version can be employed, and in a busy community any important deleted pages will normally reappear.

A good example of the power of collaborative editing on a site is the Wikipedia: a multi-language, open-content encyclopedia that is collaboratively edited by "netizens" and hosted on a wiki-based system. Wikis are suitable for exporting metadata in semi-structured format since the wiki pages are normally already stored in a database. Some wikis provide RSS newsfeeds describing changes or additions, and projects like Platypus Wiki are aimed at producing wiki pages and rich metadata for these pages.

3 The SIOC Ontology

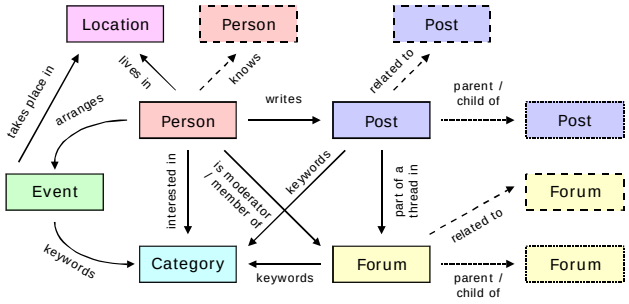


Figure 1: Relationships between the main classes needed for interlinking community sites

Most of the tools mentioned previously are already part of the HTML web. For example, being able to display messages from mailing lists in a web browser provides a comfortable way to archive posts on a website and search through the archive using HTML search technology. Linking individual posts with others is possible on the HTML level. On the Semantic Web, such content has to be made accessible by machines.

Some tools, most notably more recent tools such as weblog software, already provide export capabilities in RDF, mainly describing blogs and blog items using simple vocabularies such as RSS and Dublin Core (DC). Weblog items are linked already, but mostly only on the HTML level.

We propose an ontology that can accommodate content from all tools described previously. Most of the community tools mentioned in the previous sections store their data in semi-structured format, and more advanced sites such as bulletin boards, weblogs, and wikis use relational databases to store data. Where data is not archived such as on IRC channels, bots can provide access to utterings in chat. We describe the required ontologies to publish all this content in a semantically richer way. SIOC aims to capture all the information relevant to community sites. We have attempted to cover as broad a range of information as possible, while keeping the ontology simple enough for users to be able to navigate and browse the ontology without getting lost. An overview of the SIOC ontology is depicted in Figure 1.

There is the issue of whether one should one link and reuse some existing ontologies, or use mappings to an entirely new ontology and therefore require more intelligent applications. If a mapping is provided, there is more flexibility but algorithms need to be provided to perform the mapping and the data needs to be transformed from one format to the other. There exists mappings between existing ontologies that can be carried out using the `owl:equivalentProperty` and `owl:equivalentClass`. The problem is how to efficiently perform the various mappings.

Where applicable, we use terms from already existing ontologies, because expensive reasoning would be needed to map terms, and using terms already in existence helps to weave the Web more closely. The SIOC ontology uses terms from FOAF[7], DC[8], and RSS

[6] http://purl.org/rss/1.0/

[7] http://www.foaf-project.org/

1.0, plus newly defined terms to allow integrated access to the different formats exported by the community tools described earlier. The core concepts we identified for community sites are: person, forum, post, category, location, and event. We define new terms where needed, mainly to describe how the core concepts are related to each other.

3.1 Person

In a bulletin board-based community, profile information on each user is mainly gathered at registration through the use of fields that a user must fill in before an account can be fully activated. Required fields can include name, email address, interests, work details, and so on, which can form the basis of a FOAF file for a particular user, assuming that they have agreed to make the information publicly available. An option could be added for users that would allow the automatic creation of a FOAF file from their profile if enabled by the user. One of the useful features of a community-based bulletin board system is the "buddy list". This allows users to see when their friends are online, or to send all of their friends a private message at once. Most buddy lists are private to a particular user, but by adding an option to the bulletin board software to make the list of buddies publicly viewable, the public buddies could be used as part of a FOAF export.

The descriptions of persons in SIOC can come from various sources. The main vocabulary used is FOAF. Other formats such as the Knowledge Web person concept are mapped to `foaf:Person`.

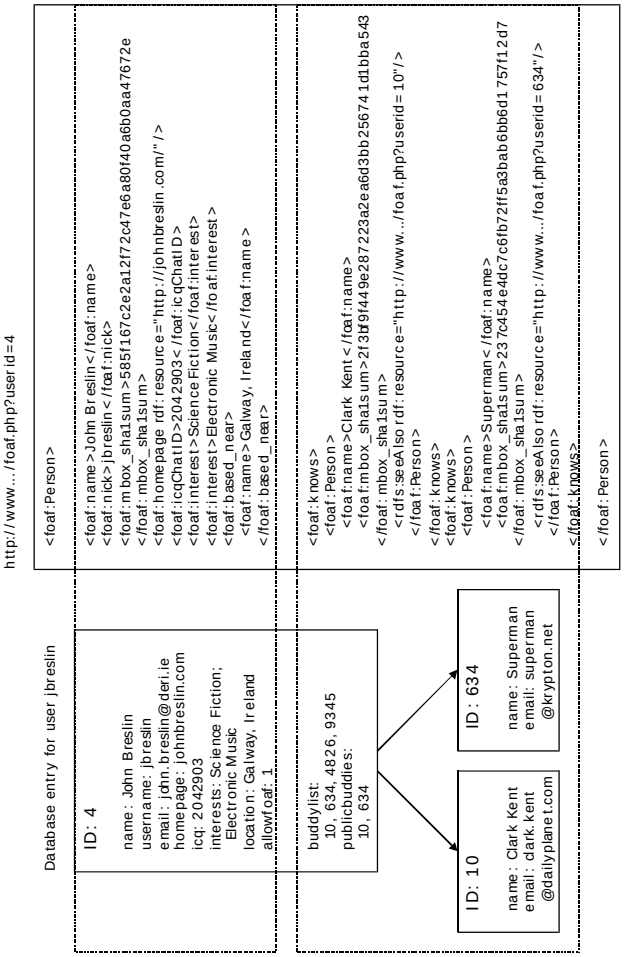


Figure 2: Example of a user profile exported to FOAF vocabulary

3.2 Post

A post in SIOC can be an article, an email, a document made available online, a piece of audio or video, a Usenet post, and so on. Although RSS is the smallest common denominator that is currently widely used to publish everything from email to weblog entries, more sophisticated descriptions to exchange posts are needed. Therefore, we define a new concept `sioc:Post` since it spans a large number of concepts from other vocabularies, such as new pages on wikis or change notifications. A special feature of `sioc:Post` is the support for threading to keep the information on what messages are direct follow-ups of other messages.

A `sioc:Post` can be either identified using a URI in case the original content is accessible via a URI, or a message ID if the original content is an email or a Usenet post that is not publicly available via IMAP4 or NTTP. Ideally, all `sioc:Post` are identified (and dereferencable) by URI. The text content of a message is enclosed in a `sioc:content` property that is CDATA, in case the origin of the message is plain text or HTML.

3.3 Event

Whereas a post can be used to announce an event, it normally lacks the structured information required by

[8] http://www.dublincore.org/

machines to identify event terms such as start time, end time or location. Some sites endeavour to link the two concepts post and event: e.g., on DBWorld, an email call for papers can be annotated with X-DBWorld headers that include information about the call and related event.

A broad number of activities in communities are centred on meetings, conferences, and other events. Currently, invitations to meetings, calls for papers for conferences, and other announcements of events are disseminated using email or are announced on web pages. Calendaring tools are used by individuals to manage their schedule. Some websites have designated events sections where events related to the organization are published in HTML. There exist vocabulary proposals to describe conferences, an RDF serialization of iCal[9], RSS events, and various others.

The event class in our schema can accommodate all these different types of information. Tools such as scheduling applications can make use of the event descriptions to schedule meetings or publish attendance lists at conferences. The main properties of events are their start and end dates where the event is over an interval, or just a date in the case where the event is an instantaneous item [Pan and Hobbs, 2004].

3.4 Location

There exists several proposals for encoding location. In SIOC, we use a combination of `geo:`[10] and `wail:`[11] to describe places. In our ontology, locations have an associated name and latitude/longitude values. Things that take place in the real world as opposed to cyberspace, such as events and people, are related to a geographic location [Michalowski *et al.*, 2004]. Location is an important class in online communities since often these communities are centred on a geographical area. Location instances can be referenced from elsewhere (e.g., from within an event) using the URI of the location.

3.5 Forum

A forum can be a bulletin board, a Usenet group, a mailing list, an IRC bot, a weblog, or some other container that can be the origin of a post. Posts are generally originating or belonging to a certain forum. The `sioc:Forum` concept can be used to describe containers that hold information about posts. Persons can also be related to forums in various roles, either as administrators, moderators, or members with reading and/or writing permissions. The specification of a forum contains a moderator which can approve and reject posts, and who generally has higher access rights than users. Forums can also have an associated list of users in the case where they are not publicly available. Forum containers can be used by people to link to, and therefore they can also express their membership in a board. A reverse link to the user should also exist for the purpose of authenticating users on a private forum. Forums generally cover a certain topic, which can be expressed by using `dc:subject`.

3.6 Category

Currently, category information is mostly keywords, based on natural language. Keywords work in closed-user groups, where all members of the group speak the same language and use the same word to denote a concept, but it is hard to ensure keywords across communities on a global scale. Processing natural language is difficult and the terms used to identify a subject have to be disambiguated or translated in different languages. Therefore, SIOC favours URIs to denote category information, although keywords are also allowed. Arbitrary URIs can be used to identify subjects, which enables better linking and gluing of items, rather than have to perform string matching. Already existing taxonomies such as WordNet[12], Omega[13], or dmoz[14] can be used to denote category information. Arbitrary URIs can be as well used to denote concepts, e.g. one can use http://semanticweb.org/ to denote the concept “Semantic Web”.

4 Data Integration

We have presented the SIOC ontology and described potential mappings to other vocabularies such as RSS. We have shown how to encode information from various information sources such as email, Usenet, bulletin boards, weblogs, and databases in RDF using the SIOC ontology. In this section, we will sketch an architecture that enables access to all information amongst interlinked community sites.

4.1 Linkage

The main benefit of using SIOC is the ability to link all sorts of entries from and among various community sites. Forums are linked to each other inside a community site and perhaps using a hierarchy or taxonomy. Links may exist already, e.g., if an email on a mailing list mentions the URI of a weblog entry, or a Usenet post mentions a web site. With SIOC, it is possible to produce leverage from links in an HTML document by making them explicit in a machine-interpretable format. Users or administrators can make connections between the various sites using the SIOC ontology, e.g., linking the newsgroup “comp.os.linux” to the mailing list “ILUG”, and vice versa. Links from people to post or forums can be drawn automatically, because the forum has information on who wrote a particular post. SIOC enables any site to make this type of information available for machine consumption.

SIOC enables the linking of community sites in a machine-interpretable format. By creating the links between all sorts of containers and sites manually, the result is a network of community sites. Connections

[9] http://www.w3.org/2000/10/swap/pim/ical
[10] http://www.w3.org/2003/01/geo/
[11] http://www.eyrie.org/~zednenem/2002/wail/
[12] http://www.cogsci.princeton.edu/~wn/
[13] http://omega.isi.edu/
[14] http://www.dmoz.org/

between community sites can be made in various ways: e.g., we can infer a connection if the same person posts to different sites, or if one forum is explicitly linked to another forum at another community site.

Having additional information encoded in SIOC enables various scenarios. The information does not always have to be added, but is there already, e.g., the authorship information of posts. As a first step to leverage of the additional information, one can imagine to publish the additional data in HTML and let the user browse that information (e.g. via a social networking component that shows the links between various users in a community or inter-community). Making the information explicit offers already increased functionality. People can browse from an email in a mailing list to a website, or from a forum post to another (threaded) forum post, or from a forum post to a weblog entry at another site.

4.2 Site

We will extend the notion of "site" here. In the context of SIOC, it is not important what data the site offers, or the nature of the site. The important thing is that the data is made in a semi-structured data format, via the SIOC ontology in our case. Our aim is to semantically enable community sites, and in the description of SIOC, we detailed a common data model for the community sites. Since the data can be expressed in RDF, we can assume that it is possible to provide access in RDF as well.

Also, it is appropriate to extend our notion of community sites. If we say "sites", we mean all sorts of community tools as mentioned in the previous section. For the purposes of discussion, we assume that all sites implement a common interface that allows access to the underlying data. If a site provides this interface, it can be integrated into a network of sites that exchange data. For that matter, it is not relevant whether the data comes from a site in a network of community sites, from a peer or node in a peer-to-peer network, or from a web service on the Internet. A site can be also used to denote a weblog installed on a users' computer or an instance of a semantic networked desktop [Decker and Frank, 2004] if a user not only want to share their opinions in text format but also their way of organising items in an ontologised form.

We assume that the underlying information of sites that implement a common interface is encoded in SIOC or in a format such that a mapping to SIOC is possible. We do not want to simply browse the data (since that is already enabled by the linkage), but rather to be able to automatically process the data (which is enabled through making the data accessible in RDF).

4.3 Data Warehousing

We assume that we can now fetch all the data and map it to the local ontology process automatically. Initial work has been performed by SECO [Harth, 2004] to show the general feasibility of mappings and integration. We have sites which offer access to their "raw" data, and there exists connections to other raw data sites, together with potential mappings into SIOC in case the other site does not offer their data in SIOC already. One solution would be to regularly pull all data from an external site into the local database (in a process called "data warehousing"). A crawler can be used to periodically fetch all the data from the external sites. The transformation of the schemas, if necessary, can be performed when importing all the data.

The data is now in one place, in a useful format (the native SIOC schema), with fast access. There are drawbacks, however. Firstly, the data is replicated and stored in two places. Such replication is common in today's web infrastructure (e.g. mail archiving systems for the Web replicate each mail in their own repositories, and web search engines often replicate the whole web into their own databases). Secondly, the data is not fresh when queries are issued, and changes in the underlying data are only propagated to the warehouse periodically when the "recrawl" is performed. It is also difficult to detect whether a flat file has been updated or not (without a file comparison where date information is unavailable).

Thirdly, there can be issues with scalability: you need large amounts of disk space and computing power to store all the collected data. Plus, all transformations to SIOC and storage methods have to be performed on every bit of data, no matter whether it is needed or not.

Performing data warehousing can fill a forum with a lot of content instantly, but bombarding the user with even more information to browse and digest may not help much.

4.4 Querying

By providing access to the data in machine-interpretable format, we gain a lot of functionality. However, replicating everything is not really an option. We can do better, by not only extending the data format from HTML to RDF, but also providing more elaborate access protocols than just performing HTTP GET on a flat file. Data made available in a semi-structured format allows for querying, so we need to extend the HTTP GET access interface to allow it to send queries. Therefore, a site can be integrated in our framework if firstly the site implements a common access interface e.g. W3C's Distributed Access Working Group (DAWG)[15], and secondly the site uses SIOC (or some vocabulary that is mapped onto SIOC). The more powerful queries will allow searches to take into account links between sites and forums, and the contents of the remote forums, enabling a coherent format in the site while integrating content from others.

What we have gained with SIOC and a common query interface is the possibility to gain uniform access to the repositories on both the data format and the access protocol level. The effort to enable every site with a DAWG-type access interface seems immense. However, similar efforts have occurred on the HTML web (e.g. GNU Mailman and MHonArc provide access to mailing lists and Google Groups provides access to

[15] http://www.w3.org/2001/sw/DataAccess/

Usenet newsgroups in HTML over HTTP). It is necessary to provide information from community data sources not only in HTML for human consumption, but also in RDF to enables automatic integration and information processing.

4.5 Virtual Integration

For use with SIOC, we propose an architecture known in database integration systems as "virtual integration" or answering queries over views. Data is fetched on demand, only when a query arrives and the data is needed. All necessary transformations (schema mapping, etc.) are carried out in this process. The query is translated, sent to all sources, and the resulting RDF is translated back into the caller's ontology.

This approach poses two challenges if we assume that data is available in the right format and deployed in a way such that the data can be queried: firstly, the distribution of queries (how far, how many hops), and secondly, performing the schema mapping between the different vocabularies efficiently. If nodes are distributed and a peer-to-peer SIOC system was to be used, some query routing would also be necessary. There is the question of how to perform this routing and when to stop (i.e., how to avoid loops in the network). The forum linkage inside SIOC makes it easier to do routing than in general-purpose peer-to-peer networks, since we have some (human-created) links that can be exploited.

5 Discussion

So far, we have described what types of tools are available, and what formats are or can be used to describe the information published by the community sites. We have detailed how to connect the various information components and provide integrated access to information relevant to a user request. Our approach has several advantages over other data integration architectures for the Semantic Web, but also has implications on how the Web infrastructure may have to evolve. We will now briefly discuss some of the other important issues related to the use of the SIOC ontology.

5.1 Annotating, Identifying Concepts

Annotation of documents in an online community can take two forms: machine and user annotation. For an existing community, some machine annotation may be necessary to begin with (using NLP or other methods) until a certain momentum has gathered and users begin to discover the advantages of annotating their posts and profiles as regards searching for information or matching other users. New communities could make user annotation primitives more visible so that the advantages of annotating documents are reinforced from the beginning.

Defining an ontology to describe the contents of the various community tools enables machine learning technologies to discover links between concepts that were previously hidden. Systems such as Gate/Geco[16] can be applied to semi-automatically detect relations between classes or extract semi-structured information out of plain text. Presenting the amount of information to the user in a meaningful way is another challenge. Users either need to employ filtering techniques or adaptive user interfaces to manage to navigate the information jungle made available by the Semantic Web.

Having the mapping between concepts carried out at a central place is not optimal, since one single person or organization has to provide all mappings and store all information centrally. On the Web, more likely to emerge is a distributed approach, where ontology editors provide mappings to other popular concepts, or third parties define mappings between concepts or ontologies. For community sites, not a central portal is needed, but a more democratic structure of equals. The peer-to-peer model offers such a model, where a priori there exists no distinction between community portals. Different community portals can distinguish themselves by providing advanced functionality, cover a more special topic, or use advanced technology.

Concepts detailing the content of posts in a community or a forum on a particular topic can be difficult to identify for a number of reasons. Firstly, there may be delays associated with identifying concepts using systems such as Gate in real-time, and secondly, there may be issues regarding conflicting concepts representing differing opinions in a single discussion forum or threaded post.

5.2 Trustworthiness

Our current model does neither include a measure of trustworthiness of an RDF statement, nor is the provenance of a statement taken into account. Digital signatures can be used to establish this web of trust, so that sensitive information such as karma in user profiles that constantly change and require a certain amount of trust in the correctness of the information can be securely distributed.

One of the problems with FOAF is that of identity. Simply by placing a FOAF file on one's site does not infer that the content relates to the owner of the site. It is possible to impersonate another person by entering their email address (whether sha1 encrypted or not) and other details such as phone number, full name, home page and so on. A useful trust aspect of the SIOC ontology is that the FOAF export from a bulletin board (where confirmation of an email address is required) is more reliable since the email address has been verified (assuming that the bulletin board URL corresponds to a trusted source of user accounts).

5.3 Object Identification

A central issue to the data integration [Levy, 1999] described previously is how to identify objects. Using URIs is especially difficult for old protocols such as email and Usenet news. Emails and Usenet posts have

[16] http://geco.semanticweb.org/

globally unique messages ids that could be used to identify them. There is still the issue of how to archive data in these formats, and dereference URIs so that the content of a message can be retrieved (e.g., to potentially dereference an email in a mailing list using IMAP4).

An open question is what URIs to use to identify things or concepts, not only for properties or classes, but also for places or category information. We have endeavoured to reuse already existing vocabularies for identifying properties and classes, but invented our own URIs when necessary. A similar problem arises when referencing to category information. Our idea is that over time certain URIs will emerge in communities that are accepted to denote specific terms. We expect URIs to emerge to describe concepts just by means of popular usage. Mappings can be provided later if people discover that two URIs are actually related or denote equivalent concepts.

Uniquely identifying each forum in a community site allows us to identify the containers for posts (whether it be a bulletin board, Usenet group or mailing list). A forum URI can potentially be a container for a group of forums, allowing the use of a structured taxonomy for classifying forums. A post URI can also be used by other posts to denote that the post is the parent in a threaded forum entry. With the SIOC ontology, a parent post or forum can exist on a separate community site from its children.

5.4 Peer-to-Peer

We mentioned that a node in the SIOC network does not need to be a community site, but can be anything that implements the common interface together with access to structured data. If we extend this thought further, you do not actually need all data to be on a site, but each user could store their data locally and connect to other users with similar data and interests directly, or connect to the site which serves as aggregator for content and facilitates meetings of new people or agents. In this case, the site acts not as meeting place and content provider, but only as meeting place.

Completely getting rid of the notion of sites introduces all sorts of problems, notably discovery of information that could be distributed over the whole globe, and searching on a world-wide scale is difficult in that it requires a lot of central resources with all the associated problems. Further, a complete central server is diametrical to the spirit of the Internet, which aims to be distributed.

5.5 Discovery

The (manual) linking of forums has one major benefit: discovery does not have to be carried out by some central authority (e.g., "show me all sites that provide this information"), since thematically related sites will link up through users providing links. By allowing manual linking, a network topology will result that can adhere to the standard network properties (power law, etc.), and your search space will be limited considerably.

The architecture is in a sense similar to web pages, where not every page is linked to every other page, but only somehow related pages. The linking aspect mainly addresses the discovery problem, and enables well-known graph operations for ranking a certain site or article to be used, since we expect the created graph to exhibit the power-law structure that is known in networks [Newman *et al.*, 2001]. The proposed architecture "bumps up" the "intelligence" of each page and allows for queries against their content.

5.6 Traffic

Another issue regarding the implementation of SIOC is a social aspect: by distributing the queries, some sites will get heavy (network) query traffic from other sites. Traffic inflicted on such sites can become a problem. Previously, these problems were resolved using a "robots.txt" file which could be used to limit access to the site under certain conditions. For the Semantic Web with its richer access capabilities, a more extensive version of the robots.txt might to be agreed to account for the scenarios laid out here.

5.7 Momentum

A hurdle in realizing the vision of interlinked community sites is the problem of wide technology adoption, i.e. how to convince people to open up their databases and provide access to their data. RDF data from social networks would be especially valuable for interlinking if they can be enticed to make use of the ontology. Some sites such as Ecademy[17] and Tribe[18] provide FOAF in flat files or RSS outputs from weblogs, but all social networking sites need to provide user profiles, discussion forums and threaded posts in RDF output for reasons of linking with other social networks and enhanced external searching capabilities. The driving forces can range from users to moderators to administrators, in tasks such as annotating posts to classifying forums to grouping user communities.

Mappings should be provided to and from SIOC and other ontologies, even if they do not offer full functionality. Such mappings can be created in a distributed way. They can come from users or site administrators that effectively act as brokers between the different ontologies.

6 Conclusion

We have described SIOC, an ontology that is needed to semantically enable and link current community sites. Making data available in machine-readable format is a requirement for constructing semantic web portals that enable access to the data published by different sources. We do not need a central authority that can provide all the services we need, but rather we start from our community site and discover related sites through the links that are added manually. As a

[17] http://www.ecademy.com/
[18] http://www.tribe.net/

consequence, community sites act as a “meeting place” for like-minded people and their software agents, where they can exchange ideas and links. A site from the SIOC standpoint is less formal and more a loose connection between people with the same interest.

The future potential for a peer-to-peer version of SIOC was proposed, where a user’s profile and posts are maintained on their own machine as opposed to a centralised site. The SIOC ontology as presented here is a step in the right direction, since the proposed structure consists of many interlinked communities that can potentially be further broken down into communities of interlinked individuals.

Acknowledgments

This work has been partially supported by DERI Líon. Thanks to Xuan Zhou, Anna Zhdanova and Knud Moeller for discussion. Schema definitions and example usage has been collected from the Web.

References

[Michalowski *et al.*, 2004] Martin Michalowski, Jose Luis Ambite, Craig A. Knoblock, Steve Minton, Snehal Thakkar, and Rattapoom Tuchinda. Retrieving and semantically integrating heterogeneous data from the web. In *IEEE Intelligent Systems*, volume 19, number 3, pages 72-79, May/June 2004.

[Decker and Frank, 2004] Stefan Decker and Martin Frank. The networked semantic desktop. In *Application Design, Development and Implementation Issues in the Semantic Web at WWW2004*, New York, May 2004.

[Harth, 2004] Andreas Harth. SECO: mediation services for semantic web data. In *IEEE Intelligent Systems*, volume 19, number 3, pages 66-71, May/June 2004.

[Hendler, 2004] Jim Hendler, Semantic web. Interview on *BBC Radio 4’s The Material World*, http://www.bbc.co.uk/radio4/science/thematerialworld_20040429.shtml, 2004.

[Lara *et al.*, 2004] Ruben Lara, Sung-Kook Han, Holger Lausen, Michael Stollberg, Ying Ding, and Dieter Fensel. An evaluation of semantic web portals. In *Proceedings of the International Conference in Applied Computing (IADIS04)*, Lisbon, Portugal 2004.

[Levy, 1999] Alon Y. Levy. Logic-based techniques in data integration. In *Proceedings of the Workshop on Logic-Based Artificial Intelligence*, Washington DC, June 1999.

[Newman *et al.*, 2001] Mark E.J. Newman, Steven H. Strogatz, and Duncan J. Watts. Random graphs with arbitrary degree distributions and their applications. In *Physical Review E, Third Series*, volume 64, parts 2, 026118, August 2001.

[O’Murchu *et al.*, 2004] I. O'Murchu, J.G. Breslin, S. Decker. Online social and business networking communities", In *Proceedings of the Workshop on the Application of Semantic Web Technologies to Web Communities at the 16th European Conference on Artificial Intelligence 2004 (ECAI 2004)*, Valencia, Spain, August 2004.

[Pan and Hobbs, 2004] Feng Pan and Jerry R. Hobbs. Time in OWL-S. In *Proceedings of the 1st Semantic Web Services Symposium*, March 2004.

[Papakonstantinou *et al.*, 1995] Yannis Papakonstantinou, Hector Garcia-Molina, and Jennifer Widom. Object exchange across heterogeneous information sources. In *Proceedings of the 11th Conference on Data Engineering*, 1995.

[Papakonstantinou *et al.*, 1996] Yannis Papakonstantinou, Serge Abiteboul, and Hector Garcia-Molina. Object fusion in mediator systems. In *Proceedings of the 22nd International Conference on Very Large Databases*, pages 413-424, 1996.

[Seaborne, 2002] Andy Seaborne. An RDF net API. In *Proceedings of the 1st International Semantic Web Conference (ISWC2002)*, pages 399-403, Sardinia, Italy, June 2002.

[Wellmann and Gulia, 1999] B. Wellman and M. Gulia. Virtual communities as communities: net surfers don’t ride alone. *Communities in Cyberspace*, pages 167-194, 1999.